融合时间衰减函数的改进协同过滤算法

殷佳莉,江智威,杨 毅,刘培培
(成都理工大学 信息科学与技术学院(网络安全学院、牛津布鲁克斯学院),四川 成都 610059)

摘 要:大数据时代数据量呈爆发式增长,为帮助人们在海量数据中获取自己所感兴趣的信息,推荐系统应运而生。协同过滤在推荐系统中应用广泛,针对传统协同过滤推荐算法数据稀疏、推荐精度较低,不能及时反映用户的兴趣度变化以及时效性不足等缺点,提出了一种融合时间衰减函数的改进协同过滤推荐算法。此算法在传统协同过滤算法的基础上综合考虑了时间因素的影响,用户兴趣会随着时间而变化,用户在短时间内感兴趣的物品具有更高的相似性,参考人类记忆遗忘特性,拟合人类记忆遗忘曲线得到时间衰减函数作为权重因子,在计算相似度和用户偏好程度时同时融入时间衰减函数对算法进行约束,提高短时间内物品相似度和用户兴趣度的权重,实现短期和长期兴趣度融合。实验结果表明,改进后的方法能在一定程度上提高传统推荐算法的精确率和召回率,验证了时间衰减函数的有效性。

关键词:推荐算法;协同过滤;人类记忆遗忘特性;时间衰减函数;兴趣偏好
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2022)04-0170-05
doi:10.3969/j.issn.1673-629X.2022.04.029

An Improved Collaborative Filtering Algorithm Incorporating Time Decay Function

YIN Jia-li,JIANG Zhi-wei,YANG Yi,LIU Pei-pei
(School of Information Science and Technology (School of Cyber Security,Oxford Brookes College),
Chengdu University of Technology,Chengdu 610059,China)

Abstract:In the era of big data,the amount of data grows explosively. To help people get the information they are interested in from the mass data,the recommendation system emerges at the historic moment. Collaborative filtering is widely used in recommendation systems. Aiming at the shortcomings of traditional collaborative filtering recommendation algorithms,such as sparse data,low precision,inability to timely reflect the changes of users' interest and lack of timeliness,we propose an improved collaborative filtering recommendation algorithm based on fusion of time decay function. Based on the traditional collaborative filtering algorithm,the proposed algorithm comprehensively takes the influence of time factor into account. Users' interest will change with time,and the items that users are interested in in a short period of time have higher similarity. By referring to the law of human memory forgetting and fitting the curve of human memory forgetting,the time decay function is obtained as the weight factor. In the calculation of similarity and user preference,the algorithm is constrained by the time decay function,so as to improve the weight of similarity and user interest in a short period of time,and realize the combination of short-term and long-term interest. Experiment shows that the improved method can improve the precision and recall rate of the traditional recommendation algorithm to a certain extent,which verifies the validity of the time decay function.

Key words:recommendation algorithm;collaborative filtering;law of forgetting human memory;time decay function;interest preference

0 引 言

随着互联网、大数据以及移动技术的飞速发展,这些技术在给予人们便利的同时数据量也与日俱增,用户要从庞大的信息中找到所需要的资源也变得越来越困难。为了解决信息过载的问题,帮助用户有效提取数据,提出了推荐系统[1]。推荐系统目前已经是一项比较成熟和成功的技术,已深入运用到了互联网产品的许多方面,如短视频平台、音乐和电影网站、电子商务、社交平台以及在线电子书等。搜索引擎和推荐系统都是解决信息过载的方法,但和搜索引擎相比它有如下优点:不需要用户主动提出搜索需求,而是系统从用户的行为日志中帮助提取用户行为信息,主动向用户做出推荐;能给用户带来新的物品体验,个性化程度高于搜索引擎;在推荐列表中会给用户提供更多选择。

收稿日期:2021-04-20 **修回日期**:2021-08-23
基金项目:国家自然科学基金青年基金(11905020)
作者简介:殷佳莉(1996-),女,硕士研究生,研究方向为数据挖掘、推荐系统;刘培培,博士,副教授,研究方向为通信信息处理、信息安全。

传统的协同过滤推荐算法一般只利用用户的历史行为信息,通过有限的数据挖掘用户的兴趣点,算法只关注用户行为,而没有将用户在与物品产生联系时的时间上下文信息加以利用,从而导致推荐的精度不高。针对传统协同过滤算法忽略了时间上下文关系的缺点,该文提出了以传统协同过滤算法为基础的改进算法。该算法充分利用了时间上下文信息,用户在不同时间下历史信息并不相同,时间越近越能反映用户当前行为信息。通过对人类记忆遗忘曲线进行拟合引入时间衰减函数,达到短期和长期兴趣度的融合对算法进行约束,强化最近时间的用户信息,优先对当前情况下感兴趣的物品进行推荐,从而提高推荐的精确率和召回率。

1　推荐算法相关理论

1.1　基于协同过滤的推荐算法

协同过滤是推荐算法中的经典,这个概念第一次被提出是在 1992 年 Xerox PARC 公司的 Tapestry 项目中[2],该项目创建的目的是让员工节约筛选垃圾邮件的时间。随后 GroupLens 网站利用其进行新闻筛选,帮助阅读者过滤大量的新闻,得到感兴趣的内容[3-4]。它不需要收集产品的有关信息,而是从用户的行为数据中过滤出有用信息进行分析和处理,从而为用户做出推荐的建议[5]。目前协同过滤已在各个推荐任务中(如图书、音乐、电影等)都有了非常广泛的应用,同时基于协同过滤的推荐算法在近几年的 Netflix 大奖赛中也多次获奖。

协同过滤根据算法机制的不同可分为:基于邻域的协同过滤和基于模型的协同过滤[6]。基于邻域的协同过滤是一种启发式的推荐算法,是推荐系统中的核心算法,具有直观、易实现、易于理解、准确率较高且无需长时间的训练过程等优点,得到了深入研究和广泛应用。因此,该文也将采用基于邻域的算法进行实验对比。此外根据计算角度的差异可将基于邻域的算法分为:基于用户的协同过滤(user-based collaborative filtering, User-CF)和基于物品的协同过滤(item-based collaborative filtering, Item-CF)[7],在实际应用中,基于物品的算法使用更加广泛,图 1 是协同过滤推荐算法的几种分类。

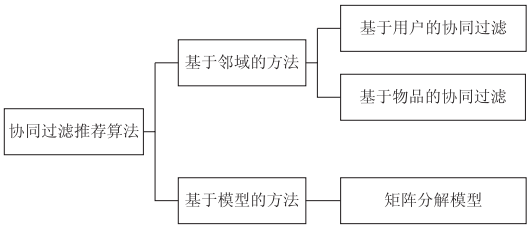

图 1　协同过滤推荐算法的分类

1.2　基于用户的协同过滤推荐算法

基于用户的协同过滤推荐算法(User-CF)以用户为研究对象,只从用户产生过行为的物品中获取特征偏好进行分析,基本原理是利用用户访问物品行为的相似性来找出相似用户,在用户与用户之间互相推荐可能感兴趣的资源,主要体现了“人以群分”的思想[8-9]。如表 1 所示,用户 A 喜欢物品[A,C],用户 B 只喜欢物品[B],用户 C 喜欢物品[A,C,D],那么可以认为用户 A 和用户 C 是具有相似品味的人,这时候就可以把用户 C 喜欢的物品 D 推荐给用户 A。

表 1　基于用户的协同过滤

	物品 A	物品 B	物品 C	物品 D
用户 A	✓		✓	推荐
用户 B		✓		
用户 C	✓		✓	✓

1.3　基于物品的协同过滤推荐算法

在实际应用中(如电子商务、视频点播),用户数量远远大于项目数量,且物品的相似度相对于用户的兴趣较稳定,由此亚马逊在 2001 年提出了基于物品的协同过滤推荐算法(Item-CF)[10]。它以物品为研究对象,在实际的使用中更加广泛。从表 2 可以看出,喜欢物品 A 的用户同时也喜欢物品 C,那么可以认为物品 A 和物品 C 是相似的,当出现了一个用户 C,若他也喜欢物品 A,那么可以认为他也会喜欢同类型的物品 C,这时便把物品 C 推荐给他。

表 2　基于物品的协同过滤

	用户 A	用户 B	用户 C
物品 A	✓		✓
物品 B		✓	
物品 C	✓		推荐

1.4　算法实现流程

基于协同过滤的推荐算法根据用户历史数据,挖掘用户与用户或物品与物品之间的相似性,并根据此预测偏好程度形成推荐列表。一般来说算法的实现步骤分为如下三步:

(1)计算相似度;

(2)根据相似度找出 K 近邻;

(3)偏好程度预测。

以基于物品的协同过滤推荐算法进行说明,第一步是利用历史信息计算物品相似度,然后找到与目标物品相似的 K 个相似物品,相似度计算是推荐系统的一个核心[11]。假设有物品 i 和物品 j, $N(i)$ 表示物品 i 有过行为的物品集合, $N(j)$ 表示物品 j 有过行为的物品集合。相似度可通过 Jaccard 相似度或余弦相似度进行计算,其中余弦相似度算法最为经典,适用于数据稀疏情况,因此该文选用余弦相似度进行计算[12],

公式如下：

Jaccard 相似度：

$$\mathrm{sim}(i,j)_{\mathrm{Jac}} = \frac{N(i) \cap N(j)}{N(i) \cup V(j)} \tag{1}$$

余弦相似度：

$$\mathrm{sim}(i,j)_{\cos} = \frac{\sum_{u \in N(i) \cap N(j)} 1}{\sqrt{|N(i)||N(j)|}} \tag{2}$$

在计算得到物品间的相似度以后，按照要求选取 K 个物品作为相似物品，利用相似度以及用户对物品的兴趣度得到用户对未知物品的偏好程度，计算用户 u 对物品 i 的偏好程度的公式如下：

$$p(u,i) = \sum_{j \in S(i,K) \cap N(u)} \mathrm{sim}(i,j)_{\cos} r_{uj} \tag{3}$$

其中，$S(i,K)$ 表示和项目 i 的 K 近邻，$N(u)$ 是用户 u 有过正反馈行为的项目集合，$\mathrm{sim}(i,j)_{\cos}$ 是物品 i 和物品 j 的相似度，r_{uj} 表示用户 u 对物品 j 的兴趣度，若用户 u 对物品 j 有过行为，则可以令 $r_{uj}=1$，否则 $r_{uj}=0$。得到偏好程度值后按从小到大排列，选出 N 个物品作为待推荐物品，这种方法也称为 Top-N 推荐。

2 融合时间衰减函数的推荐

2.1 时间上下文

传统的协同过滤推荐算法一般只利用用户的历史行为记录和评分数据挖掘用户的兴趣偏好，进而向用户推荐感兴趣的物品，而忽略了用户发生行为时所处的环境，这里的环境也指用户所处的上下文信息，如：时间、地点、心情等。

其中时间信息是上下文信息中的一个重要因素，时间信息对推荐系统的影响体现在：用户兴趣是变化的；物品有生命周期的；季节效应[13]。例如，用户在小学阶段对动画的兴趣更高，但随着年龄增长而转向电视剧。有些项目生命周期比较短，如新闻。招聘信息、促销活动等，一旦时间超过项目生命周期，再进行推荐就失去了意义。此外特定项目还会受到季节效应影响，如夏天穿短袖，冬天穿羽绒服等。即考虑到人的偏好行为特性是随着时间而变换的，在不同时间段，人的行为特征是不一样的。此外，比起其他上下文因素，时间上下文信息是最容易获取的情境信息，可以通过系统时钟、事务时间戳等方法隐式地获取时间[14]。因此，综合考虑时间对用户兴趣度和物品相似度的变化影响，将时间上下文信息融入传统协同过滤推荐算法中，是该文研究的出发点。

2.2 人类记忆遗忘规律

在心理学研究中，人类的记忆可以区分为短时记忆和长时记忆，短时记忆容量有限，如果不在一定时间内回顾，这些短时记忆很快就会被遗忘。人类的记忆是有限的，如今天学的知识会清晰记得，但是在昨天学的知识可能会有些模糊，时间更久一点或许就完全遗忘。德国著名心理学家赫尔曼·艾宾浩斯，通过实验研究人类记忆遗忘变化情况，并通过实验结果总结出人类记忆遗忘规律[15]，如表 3。从该表格可以看出人类的记忆在第一天内大部分都会被遗忘，在此之后遗忘的速度开始变得平缓，最终只保留下一小部分。

表 3 艾宾浩斯遗忘规律

时间间隔	记忆保留/%
刚记忆完	100
20 分钟后	58.2
1 小时后	44.2
8-9 小时后	35.8
1 天后	33.7
2 天后	27.8
6 天后	25.4
一个月后	21.1

将实验结果绘制成艾宾浩斯遗忘曲线，如图 2 所示。可以看出，人类对事物的遗忘速度并不是一直不变的，刚开始的遗忘速度很快，随着时间的推移遗忘速度会慢慢变得平缓，这是一个“先快后慢”的过程。

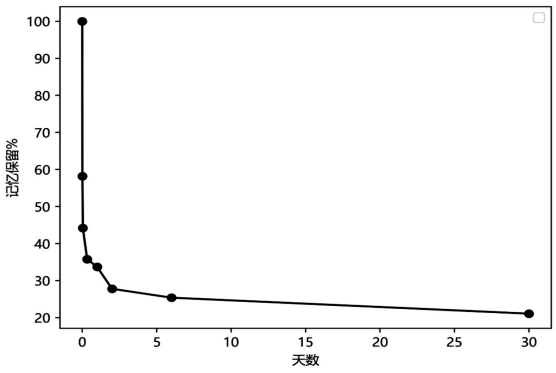

图 2 艾宾浩斯遗忘曲线

用户的兴趣变化与人类记忆特性相似，通过类比人类记忆遗忘特性，可以认为用户对物品的兴趣度会随着记忆的遗忘而逐渐衰退，用户可能会对短时间内关注的物品具有更高的兴趣度，用户的短期行为应该受到更高的关注。在推荐系统中通过类比人类记忆遗忘特性，时间越近越能体现用户的兴趣度，物品间的相似度也越高，短期兴趣在推荐中的权重也越高。

2.3 时间衰减函数

通过类比人类记忆遗忘规律拟合时间衰减函数，用时间衰减函数来表示用户对物品兴趣的权重，用户对物品产生行为的时间越长，对用户现在的兴趣影响就越小。衰减函数的拟合方式可以有线性、指数和对数形式[16]，$f(|T-t|)$ 为线性时间衰减函数，其表达式如下所示：

$$f(|T-t|)=\frac{1}{1+\alpha|T-t|} \tag{4}$$

其中,α 代表时间衰减因子,α 决定着 $f(|T-t|)$ 的衰减速率,通过调整取值即可模拟记忆遗忘曲线,如果用户兴趣变化越快则衰减速率越快,α 的值要大一些,在不同系统中应根据实际情况设定;T 表示当前时间;t 表示产生行为的时间;$|T-t|$ 表示时间差,$f(|T-t|)$ 随着时间差的增长而减小,即取值范围为 $(0,1)$。

2.4　融合时间衰减函数的推荐

融合时间衰减函数的推荐算法需要提高用户近期行为的权重,把时间衰减函数作为权重因子对用户或物品相似度进行约束,用户近期行为相比用户以前的行为更能体现用户现在的兴趣。考虑时间的影响因素,在传统的协同过滤相似度计算公式上进行改进,优化后的余弦相似度计算公式如下:

$$\mathrm{Tsim}(i,j)_{\cos}=\frac{\sum_{u\in N(i)\cap N(j)} f(|T-t|)}{\sqrt{|N(i)||N(j)|}} \tag{5}$$

除了在相似度计算时融入时间衰减函数外,也应该考虑时间信息对偏好程度的影响,用户的近期行为相比用户远期行为更能体现用户当前的兴趣。因此,在预测用户当前的兴趣时,应该将用户近期反馈项目的权重增大,优先推荐与用户近期喜欢或购买过的项目相似的项目,在得到用户对项目产生行为的时间后,可计算用户的兴趣偏好,修正后的偏好程度见公式(6),其中 β 为时间衰减因子,需要根据不同的数据集调整值的大小。

$$\mathrm{Tp}(u,i)=\sum_{j\in S(i,K)\cap N(u)} \mathrm{Tsim}(i,j)_{\cos}\frac{r_{uj}}{1+\beta|T-t|} \tag{6}$$

3　实验和结果分析

3.1　实验设置

实验方法:实验将采用离线实验对比融合时间衰减函数的协同过滤推荐算法与传统协同过滤推荐算法的离线性能。

实验数据集:选择 delicious-2k 数据集,它包含了 1 867 名用户,105 000 个书签和 69 226 个网址信息,因为网页由 URL 标识,因此可以根据域名将网页分成不同的类别。从中获取域名为“www. nytimes. com”的数据集进行实验。

3.2　算法评估指标

Top-N 推荐一般通过精确率与召回率进行衡量。其中 U 表示用户集,$R(u)$ 是根据用户在训练集上的行为给用户做出的推荐列表,$T(u)$ 表示用户在测试集上的行为列表。精确率的意义在于所预测的推荐列表中有多少是用户真正感兴趣的,预测列表的精确率可以直接反映推荐的好坏[17],精确率的定义为:

$$\mathrm{Precision}=\frac{\sum_{u\in U}|R(u)\cap T_{(u)}|}{\sum_{u\in U}|R(u)|} \tag{7}$$

召回率表示用户真正感兴趣的列表中有多少是被推荐算法准确预测出来的,即真实列表的召回率,召回率的定义为:

$$\mathrm{Recall}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{8}$$

3.3　实验过程

Nytimes 数据集中包含了 443 名用户的行为数据,对数据集进行提取整理得到 (u,i,t) 的三元组。对每个用户的行为按照时间戳由早到晚排序,时间越近排序越靠前,由于一名用户对书签有多个行为数据也就有多个时间戳,将用户最后一个时间戳的行为作为测试集,测试集包含了 443 名用户的 443 条行为数据;将 443 名用户在最后一个时间戳之前的所有行为记录作为训练集。

在基于物品的协同过滤算法实验中,近邻个数设置为 10,计算物品间的相似度时采取余弦相似度;在实现融合时间衰减函数的协同过滤算法实验中,近邻个数设置为 10,分别使用改进后的余弦相似度和兴趣偏好程度进行计算。

改进后的算法根据训练集学习用户兴趣模型,给每个用户推荐 N 个物品,该文将选取不同 $N(1,2,\cdots,10)$ 进行 10 次实验,得到融合时间上下文的协同过滤算法与未融合时间上下文的协同过滤算法的精确率和召回率,并将实验结果绘制成表格和折线图进行结果对比。

3.4　实验结果

利用优化过的融合时间衰减函数的协同过滤算法在进行离线测试时,采用 TOP-N 列表推荐,最终的评估指标选择精确率和召回率。在数据集上分别比较融合了时间上下文的协同过滤与未融合时间因子的协同过滤的实验结果,如表 4 和表 5 所示。由这两个表格可知:在相同推荐长度下,改进后的算法的准确率和召回率均优于传统算法;将精确率和召回率绘制成折线图分别如图 3、图 4 所示:随着推荐长度逐渐增加,推荐精确率逐渐下降,而召回率有所提高,但改进后的算法的效果比传统算法更优。

通过图表可以得出结论:融合时间衰减函数的协同过滤推荐算法和传统协同过滤推荐算法相比,能在一定程度上提高推荐的精确率和召回率。

表 4 NYtimes 数据集不同推荐长度精确率对比 %

方法	1	2	3	4	5	6	7	8	9	10
Item-CF	5.21	2.72	2.07	1.75	1.86	1.99	1.85	2.00	1.89	2.26
TItem-CF	7.29	3.8	3.32	2.81	2.48	2.56	2.38	2.24	2.13	2.26

表 5 NYtimes 数据集不同推荐长度召回率对比 %

方法	1	2	3	4	5	6	7	8	9	10
Item-CF	1.13	1.13	1.13	1.13	1.35	1.58	1.58	1.81	1.81	2.26
TItem-CF	1.58	1.58	1.81	1.81	1.81	2.03	2.03	2.03	2.03	2.26

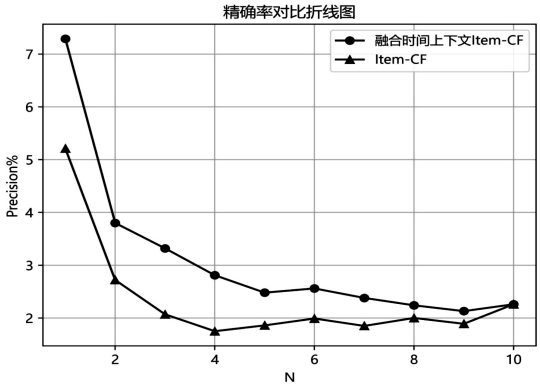

图 3 NYTimes 数据集的精确率曲线

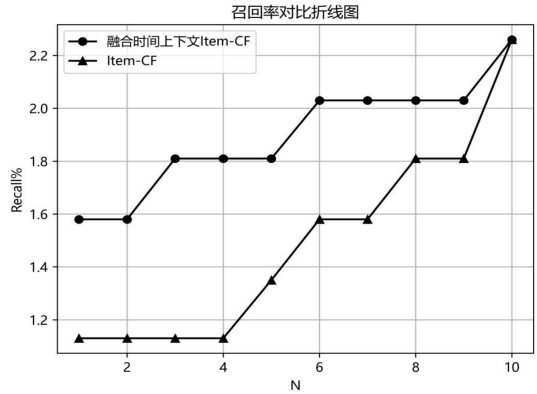

图 4 NYTimes 数据集的召回率曲线

4 结束语

将时间上下文信息引入传统协同过滤中以获得更好的推荐精度,由此提出了一种融合时间衰减函数的改进协同过滤推荐算法。参考人类记忆遗忘特性曲线,拟合时间衰减函数与传统推荐算法结合,建立与时间相关的基于物品的协同过滤推荐算法,从而做出推荐。通过实验测试,验证了所提出的融合时间衰减函数的协同过滤推荐算法相比传统协同过滤算法能在一定程度上提高推荐的精确率和召回率,验证了时间衰减函数的有效性。

参考文献:

[1] 游 聪. 基于时间上下文的个性化推荐技术研究[D]. 成都:电子科技大学,2016.

[2] 孙光福,吴 乐,刘 淇,等. 基于时序行为的协同过滤推荐算法[J]. 软件学报,2013,24(11):2721-2733.

[3] ZHANG Wei,WANG Jianyong. Location and time aware social collaborative retrieval for new successive point-of-interest recommendation[C]//Proceedings of the 24th ACM international on conference on information and knowledge management. Melbourne,Australia:ACM,2015:1221-1230.

[4] RESNICK P,IACOVOU N,SUCHAK M,et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. Chapel Hill,NC:ACM,1994:175-186.

[5] LIU Q,CHEN E H,XIONG H. Enhancing collaborative filtering by user interest expansion via personalized ranking[J]. IEEE Transactions on Systems Man Cybernetics,2012,42(1):218-233.

[6] DESROSIERS C,KARYPIS G. A comprehensive survey of neighborhood-based recommendation methods[M]//Recommender systems handbook. Berlin: Springer,2011:107-144.

[7] 刘 云,王 颖,亓国涛,等. 时间上下文的协同过滤 Top-N 推荐算法[J]. 计算机技术与发展,2017,27(7):79-82.

[8] 何 蓉. 基于卷积神经网络的音乐推荐系统[D]. 南京:南京邮电大学,2019.

[9] 张硕硕. 基于上下文和标签相关性的推荐算法研究[D]. 徐州:中国矿业大学,2019.

[10] 翁小兰,王志坚. 协同过滤推荐算法研究进展[J]. 计算机工程与应用,2018,54(1):25-31.

[11] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[12] 梁思怡,彭星亮,秦 斌,等. 时间上下文优化的协同过滤图书推荐[J]. 图书馆论坛,2021,41(3):113-121.

[13] 项 亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012:122.

[14] 刘宏志. 推荐系统[M]. 北京:机械工业出版社,2020:122.

[15] 吕海波. 基于时间上下文的个性化电影推荐算法的研究与应用[D]. 徐州:中国矿业大学,2020.

[16] DING Yi,LI Xue. Time weight collaborative filtering[C]//Proceedings of the 14th ACM international conference on information and knowledge management. Bremen,Germany:ACM,2005:485-492.

[17] 刘 攀,陈敏刚. 个性化推荐系统评估[J]. 南昌大学学报:理科版,2016,40(2):143-150.